

2024上海图书馆开放数据竞赛巡讲·四川大学

计算伦理与数字道德

——AGI时代的人文主义和数字人文



刘炜 上海图书馆上海科技情报所

wliu@libnet.sh.cn





保障可信的/负责任AI

1. 信息过载与信息茧房

2. 虚假信息

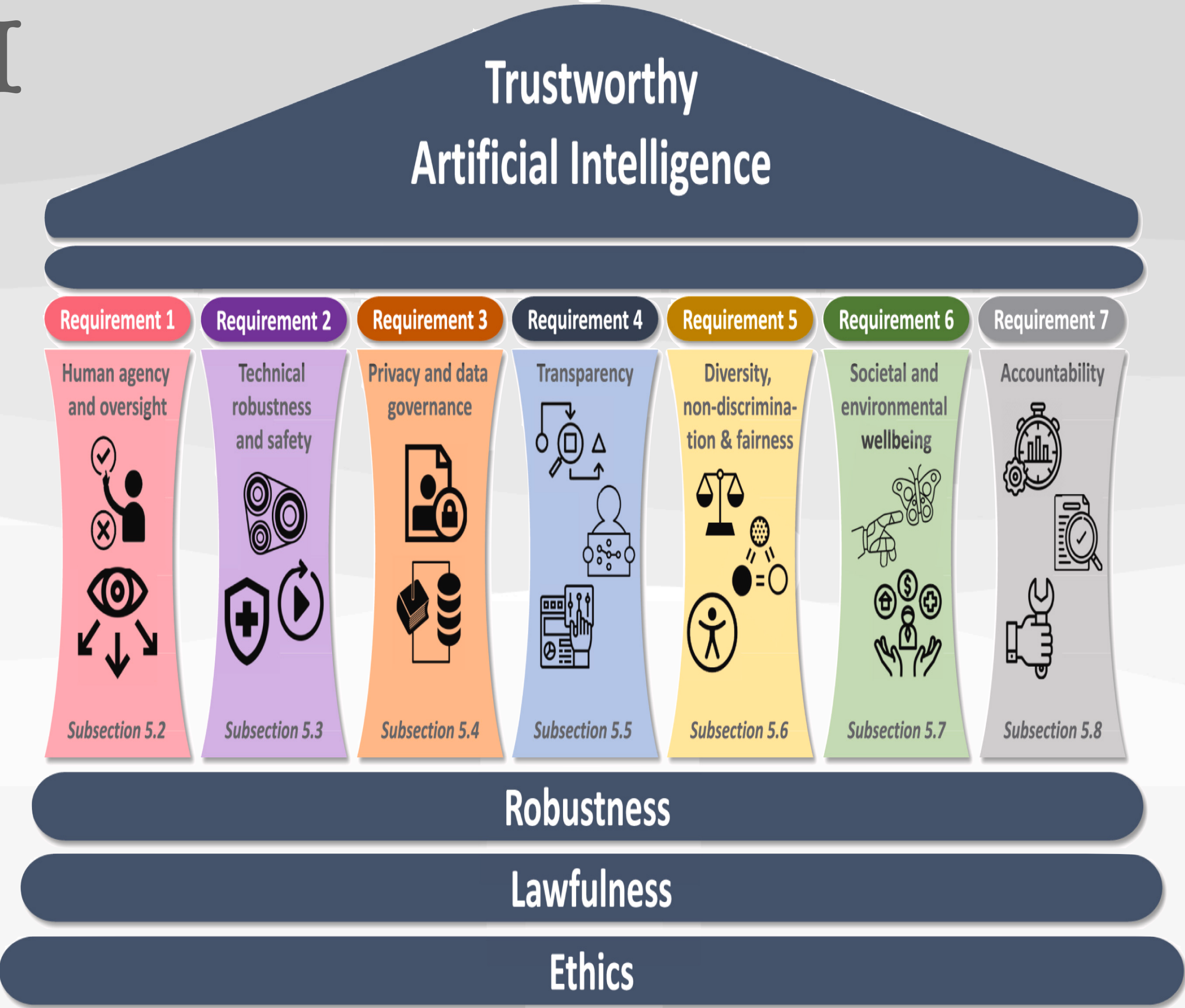
3. 技术素养与失业问题

4. 误用滥用与责任边界
5. 侵犯隐私与信息泄漏

6. 侵犯版权与诱导犯罪

7. 军事应用与生物威胁

8. 意识觉醒与情感欺骗



计算伦理与数字道德

- 计算伦理：伦理是指对行为的规范和原则的系统研究，是价值观的组成部份。计算伦理关注和评估计算机技术对社会、个人及道德标准的影响，它涉及如何使用计算机技术的行为规范，包括数据隐私、安全性、公平使用和知识产权等方面，也包括广泛应用了计算技术之后，尤其是产生了各类高度发达的计算机智能体之后，伦理规范的主体性和差异性问题的，探讨伦理的边界和机器伦理问题。
- 数字道德：道德是指个人或社会普遍接受的关于善恶、正义和义务的信念和行为规范，通常是基于文化、宗教和社会习俗的具体准则和规范。数字道德涵盖了所有数字技术的使用对社会伦理和个人行为的影响，甚至对人类的约束映射和延伸至高级智能体。数字道德探讨各类实体在数字世界中如何维持道德行为，包括社交媒体的伦理、数字隐私、元宇宙中的行为规范等等。
- 人文主义（Humanism）是一种哲学思想和文化思潮，强调人类价值、尊严和潜力，注重个人的理性、自由和道德责任。人文主义起源于文艺复兴时期，但其思想根源可以追溯到古希腊和古罗马的哲学传统。由于计算伦理与数字道德的外延扩大到了人类之外，人文主义的边界是否有必要扩展，是一个需要深入探讨的问题。

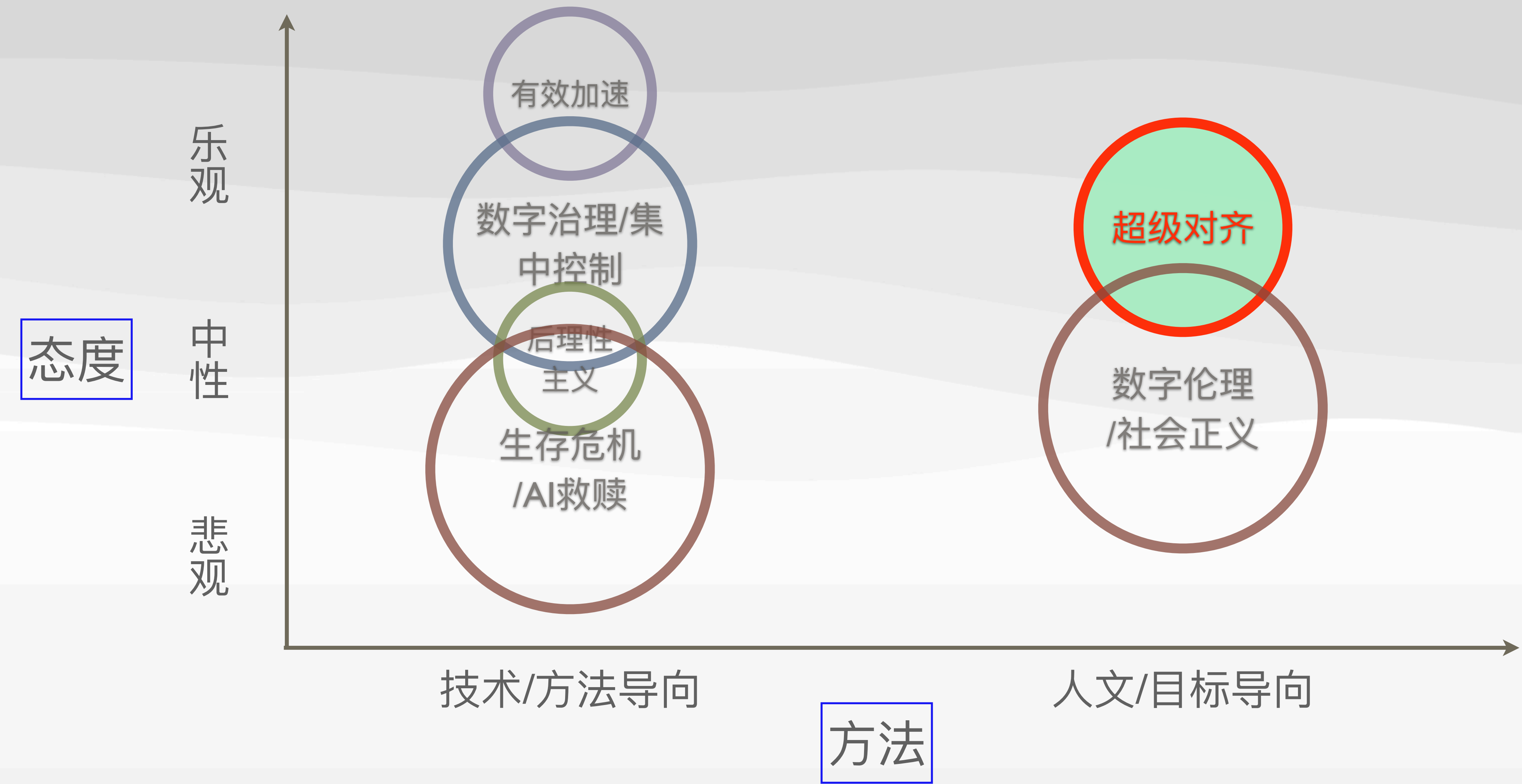
什么是人文主义：拉里佩奇与马斯克的分歧



E/ACC与SUPER LOVE ALIGNMENT



有效加速主义与超级对齐主义

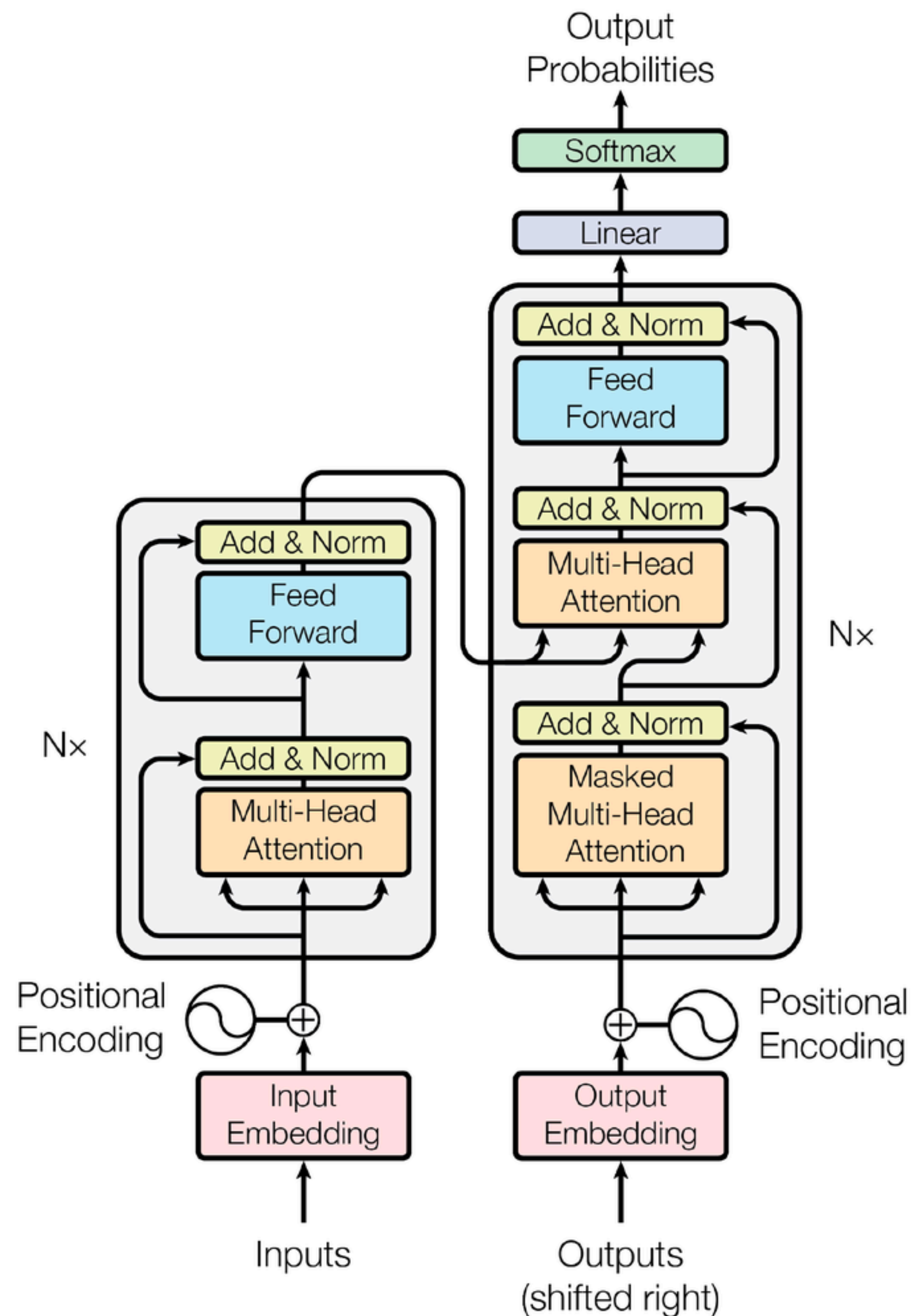


AGI之梦

- AGI（通用人工智能：Artificial General Intelligence）是指一种具备人类智能水平的机器，能够理解、学习和应用各种不同领域的知识。与当前专注于解决特定问题（如图像识别、自然语言处理或棋类游戏）的人工智能不同，AGI能够执行“任何”智能任务，包括任何领域的推理、规划、学习、交流，以及感知和与物理世界互动的能力。
- AGI四要素：1.具有通用智能，能够跨领域学习和工作；2.能够自我学习和适应环境；3.能够理解复杂/抽象概念并进行逻辑推理；4.甚至具有情感和意识！
- 目前的大语言模型还不是AGI，但在语言理解和多模态方面初步具备了泛化/涌现和推理能力，是目前最有可能发展成AGI的模型技术。但通过堆数据和堆算力是否能到达AGI还有争论，但普遍认为目前还未边际效应递减，但数据和电能都几近枯竭。

LLM如何被点化？

- 大型语言模型（LLM）是基于海量自然语言数据进行预训练而得到的超大型深度学习模型，参数通常从数十亿到超千亿。底层基于Transformer深度神经网络，由具有自注意力功能的编码器和解码器组成，但GPT只采用了解码器，从一系列文本中提取含义，并能够理解其中的单词和短语之间的关系。
- 用同样方法对海量图片、音频、视频等多媒体信息结合语言数据进行预训练和指令微调的超大型深度学习模型也是大语言模型的一种发展，通常称为多模态大模型。



大模型所具备的AGI特征

■ “泛化”能力

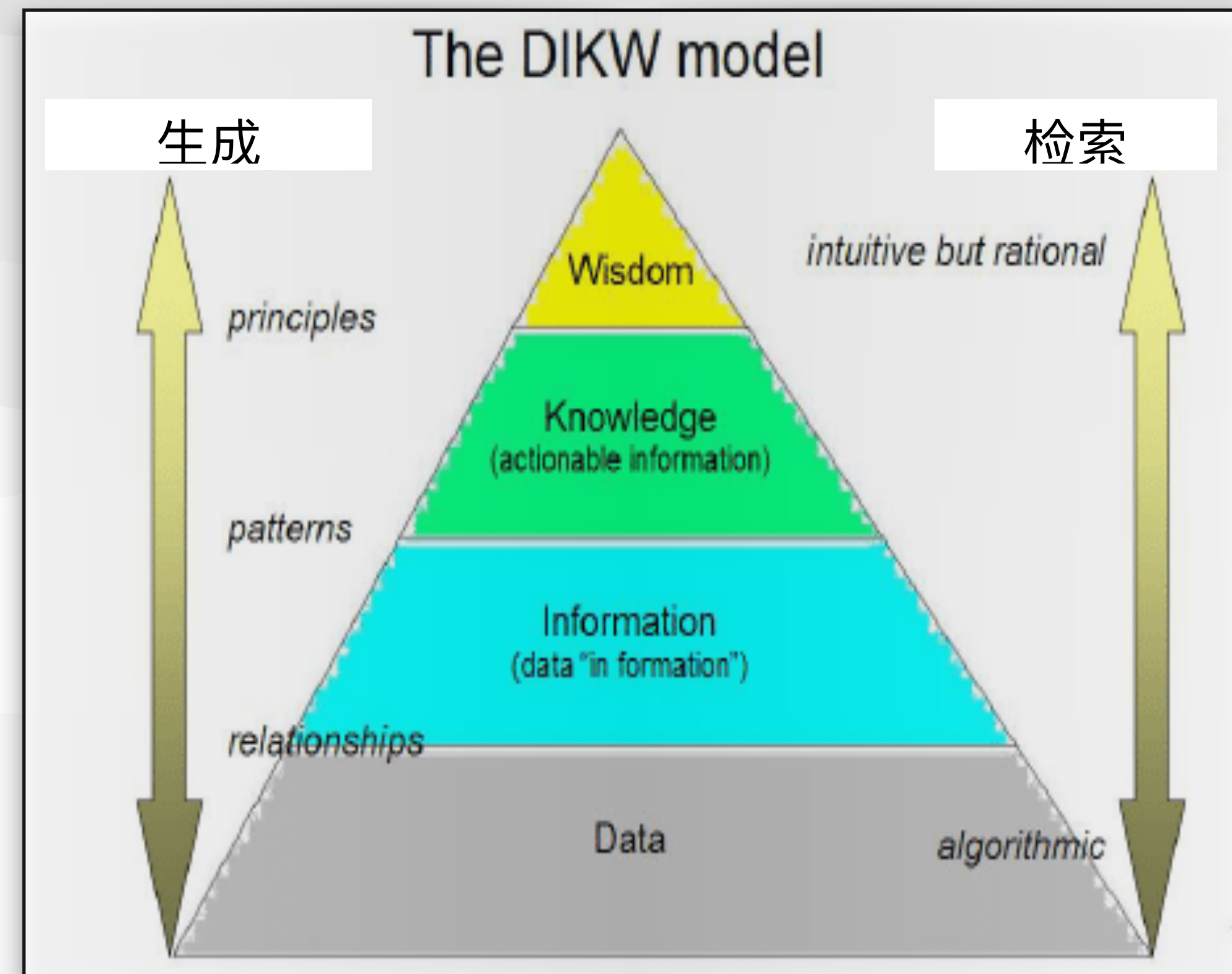
- Perplexity（困惑度）评价：用于评估语言模型在未见过的数据上的预测能力。困惑度越低表示模型在未见过的数据上表现越好。
- 语言模型的交叉验证：将数据集分为训练集、验证集和测试集，通过在验证集和测试集上的性能来评估模型的泛化能力。
- 零样本任务（Zero-shot Task）能力：在模型未见过的任务上进行评估，例如对模型提出一些与训练数据不相关的问题，评估其在这些任务上的表现能力。

■ 推理能力

- 自然语言推理（NLI）任务：在给定前提和假设下，能否正确推断出假设的真假。
- 文本蕴含任务：在给定前提和假设下，能否判断假设是否可以从前提中推导出。
- 逻辑填空任务：要求模型填写一些语句中的空白，使得整个语句逻辑上合理。
- 逻辑推理任务：要求模型根据一些逻辑规则进行推理，例如判断一些命题是否成立或给出逻辑结论。

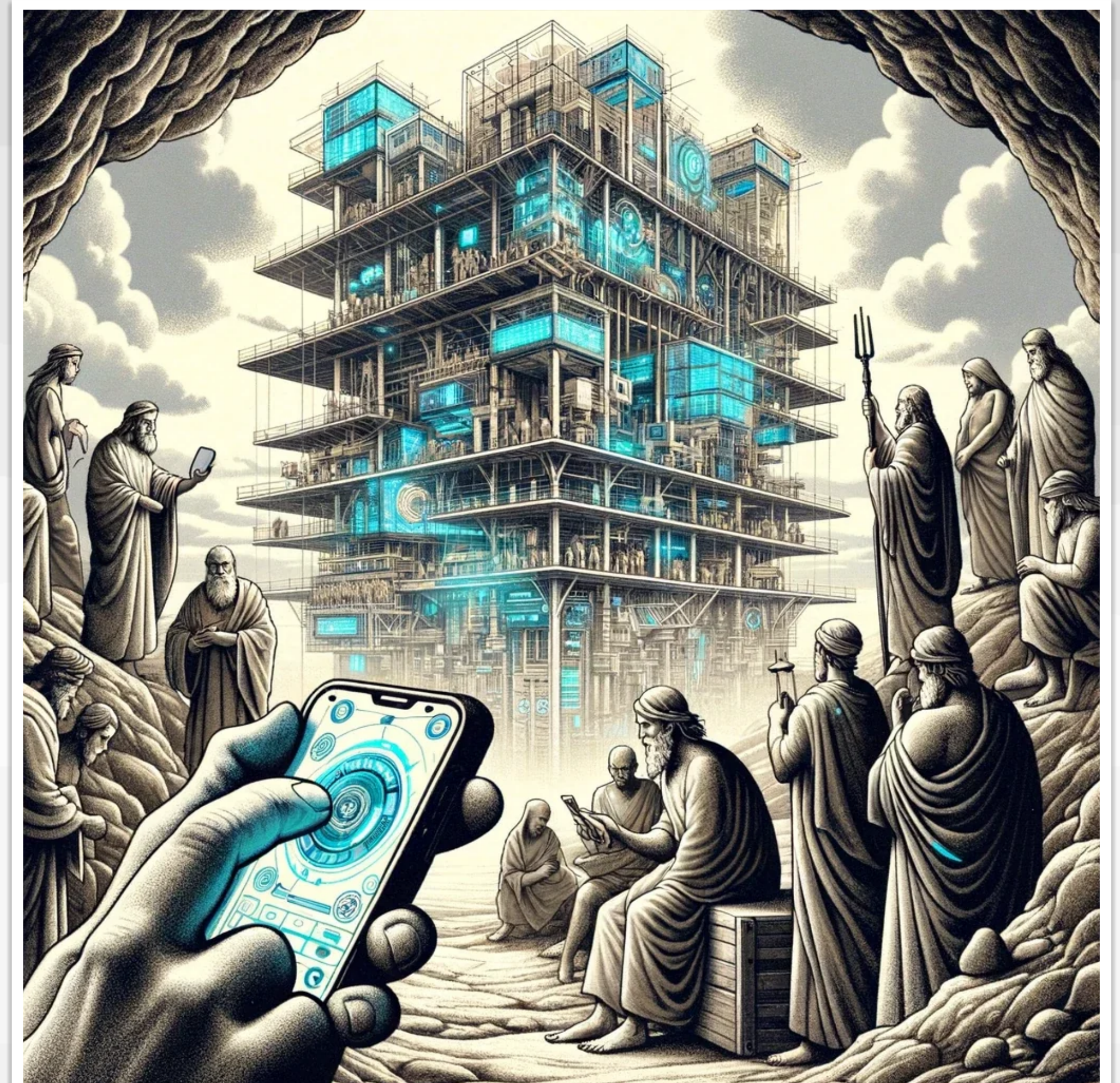
大模型的颠覆性

- LLM是海量知识在深度神经网络中的压缩形态，可以认为是智慧的一种编码存储形式。
- LLM是基于词元（token）而不是符号的：词元是一种张量，是语义相似性的度量而不是符号匹配，因此它可以直接告诉答案而不是符号的排列组合。
- 知识是随时生成的而无需预先存储的：存储知识只是AI的Bootloader，具有具身学习能力的智能体无需存储知识，只需要参数权重存储的智能即可以随时产生知识。
- LLM应用以端到端模型为最高形态，端到端是指只要给定数据就能得到智慧。
- 人类作为知识链的起点，其知识生产虽然节能，但却是极其原始而粗糙的。LLM一旦形成便不再需要人类的帮助，可以通过自我学习（自己创造数据进行学习）而得到，并在应用中不断迭代（数据飞轮）。
- LLM短期赋能传统的知识工作，长期将会颠覆整个知识产业模式。



捍卫人文主义

图书馆是黑暗森林中的灯塔，
图书馆员是AI时代的领航员。



2024上海图书馆开放数据竞赛巡讲·四川大学

谢谢!



刘炜 上海图书馆上海科技情报所

kevenlw@gmail.com